The Cost of Digital Preservation

Project Report v. 1.0

A project funded by the Danish Ministry of Culture:
Committee for Digital Preservation

Ulla Bøgvad Kejser (The Royal Library), Anders Bo Nielsen, and Alex Thirifays (Danish National Archives)

November 2009

Contents

Executive Summary

Background

The project was set up as a result of a proposal to the Ministry of Culture's "Committee for Digital Preservation" from The Royal Library, (KB), the State and University Library (SB) and the Danish National Archives (DNA) for the development of a cost model for digital materials. The proposal and thus the project were inspired by the British LIFE project[1], which in 2006 published a cost model.

Goal

The project has as its goal to set up a generic model for the calculation of costs related to preservation of digital materials at ALM institutions.

Participants in the project

The participants in the project are two staff members from DNA and one from KB who divide the nine full-time equivalent months financed by the Ministry of Culture through the committee.

Major activities

The project consists of a series of major activities which are listed below, and its products consist of a series of reports, of which this is one, and appendices. The activity numbers refer to the project initiation document of the project.

5.3.1 Investigation of the LIFE Project and its follow-up LIFE 2

5.3.2 Investigation of related cost models

5.3.3 Development of the cost model on a theoretical basis

5.3.4 Examination of Danish studies and experience and adjustments of the model to these

5.3.5 Collection of information on holdings and the expected growth in digital materials

5.3.6 Application of the model to Danish conditions (based on the collected information)

The project has focused on the preparation of:

- A general cost methodology for digital preservation, including a function model based on the OAIS Reference Model[2] and a set of generally accepted accounting principles

[1] LIFE Project website: http://www.life.ac.uk

[2] Reference Model for an Open Archival Information System (OAIS)(CCSDS, 2002)

- A cost model for functional preservation by the preservation strategy digital migration
- A set of formulas which can transform input to cost data
- A tool set (including documentation of the calculations), which makes it possible for the user to carry out automated cost calculations in relation to functional preservation

Investigation of LIFE and related cost models

Investigation of the project LIFE and its follow-up LIFE 2, as well as related cost models, showed that there is no completely developed model, but several from which inspiration can be gained, among others, the LIFE Costing Model, KRDS1[3], NASA CET[4] and Testbed[5]. We had expected that we could merely expand on the LIFE model.

In the development of our cost model we have been inspired by the above mentioned models, our own experience, as well as the literature on digital preservation and cost models.

We have also investigated what Danish experience and studies exist, and can conclude that this material is exceedingly scarce.

Development of the model

Rather than an expansion of an existing model, the project has had to develop its own independent model (called Cost Model for Digital Preservation, abbreviated CMDP) based on the OAIS standard. OAIS has the best foundation since it describes all required areas and uses a functional subdivision. The model is as yet only operationalised in relation to functional preservation based on the migration strategy.

In addition to the description of the model itself in this report, it has also been set up as a user-oriented spreadsheet.

Cost critical activities

The method used for estimating cost of digital preservation, has been to breakdown the OAIS functions into cost critical activities, and continue to subdivide until measurable components were identified. These components were thereafter operationalised through formula and cost factors. An example is the cost of understanding a particular preservation format, added up as person hours, (and thence as wages), based on the time it takes to read documentation, which is based on its number of pages, difficulty and quality.

---

[3] Keeping Research Data Safe (KRDS) Project website: http://www.beagrie.com/jisc.php
[4] NASA Cost Estimation Tool (CET): http://opensource.gsfc.nasa.gov/projects/CET/CET.php
[5] Digital Preservation Testbed: www.digitaleduurzaamheid.nl (the link is broken, November 2009)

Testing on cost data

The model has been simulated on cost data from two digital migration projects (cases) carried out by DNA between 2005 and 2009. It should be noted that the same projects are an important part of our experience, and thus the model has not yet been tested on completely independent data.

The test showed that the model assumes compliance with a certain norm. The model underestimated the large project management costs for the one very extensive migration and the large error correction costs due to poor compliance with preservation standards. In the other, very small scale project the model overestimated the costs of system development and design of new data structures. The precision of the model is thus highly dependent on being used on a normal situation in relation to OAIS.

Success criteria and fulfillment

The success criteria of the project were thus described in the project initiation document:

- A closer investigation of the project LIFE and its follow-up 2 as well as related cost models
- Development of a cost model which is tested in detail on Danish data and experiences
- Use of the cost model on the estimated amounts of digital material which must be preserved for the state archives, libraries and museums in the next ten years

The two first items have been completed, though not for all the functional entities in OAIS for the second item. The last item remains unfinished, primarily because within the time limits it has not been possible to obtain and standardise the desired information from the cultural heritage institutions.

In our opinion, the model is a good foundation for further development and operationalising of the remaining function areas, both with regard to the premises, principles, methods, formulas and the user interface of the model.

1 Introduction

Digital preservation is a relatively new discipline at archives, libraries and museums (ALM institutions). Even though the DNA have received digital archival materials since 1968 and KB has received digital materials through legal deposit since 1987, not until the last 5-10 years have the institutions established actual organisations and systems for long term digital preservation. In this report we denote such systems OAIS archives[6]. Because digital preservation is complex, technological development is renewing and systems and methods constantly change the institutions have not, as yet, much experience with how to assess the costs of digital preservation or what it costs over time. The Ministry of Culture and the ALM institutions, however, need this knowledge in order to be able to set priorities and plan their activities. Furthermore, understanding the nature of digital preservation cost is a prerequisite for increasing the overall efficiency, and thus achieving cost reduction as well as first quality for preservation of cultural heritage materials. This was the background for the project proposal.

Digital preservation designates the methods and systems which are needed to assure access to digital materials over time. Digital preservation may be divided into bit preservation, which must ensure that the bits remain intact and readable, and functional preservation, which must ensure that the bits remain comprehensible. However, the concept digital preservation is also used in a broader context for the whole life cycle process from the production of the digital materials, acquisition by the OAIS archive, and long term preservation, until they are made available to the users. To avoid confusion regarding the concept of digital preservation we use in this report the term digital preservation for the whole life cycle and bit preservation and functional preservation respectively for the specific preservation functions.

The project uses the term cost model to denote a toolkit for costing consisting of a well-defined function model and explicit accounting principles (Sanett, 2002). Furthermore, it includes a description of formulas and factors used to operationalise the model.

1.1 The goal of the project

The goal of the project was to produce a model for assessing the costs related to the preservation of digital materials at the state's ALM institutions. It was set up as a result of a proposal from KB, SB and DNA to the "Committee for Digital Preservation" of the Ministry of Culture. As the following list of the main activities of the project makes clear, the proposal was inspired by The British LIFE project, which in 2006 published a cost model:

5.3.1 Investigation of the LIFE Project and its follow-up LIFE 2

5.3.2 Investigation of related cost models

[6] The standard describes an OAIS archive as "consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a designated community".

5.3.3 Development of the cost model on a theoretical basis

5.3.4 Examination of Danish studies and experience and adjustments of the model to these

5.3.5 Collection of information on holdings and the expected growth in digital materials

5.3.6 Application of the model to Danish conditions (based on the collected information)

Furthermore, the project wished to obtain an impartial evaluation of the project and has therefore asked the consultancy Charles Beagrie Limited[7] to carry out this task.

This project report is organised by a description of the results of the main activities. It is intended for administrators, preservation specialists, and others who are interested in the costs of digital preservation and how to assess these costs.

Part of the results has been presented in an article at the iPRES 2009 Conference[8], where the article was also presented as a lecture (Appendix 1).

1.2 Goals for the cost model

The target group for the cost model is the ALM institutions, which adhere to the principles for long term preservation, as described in the OAIS standard. This is not intended as a precise definition, but an indication of, that we assume that the users of the model are institutions which work professionally with digital preservation and in a normally efficient way. Envisioned users are practitioners and experts in digital preservation.The goal is that the cost model should be able to deal with all the types of materials that the ALM institutions must preserve and calculate costs for relevant preservation strategies such as migration and emulation.

The cost model must be based on a well defined breakdown of all the activities that are encompassed by digital preservation. This activity based approach must cover both the costs of the archive systems themselves and the costs related to the flow of data through the OAIS archive. Furthermore, it must be able to take into account outside factors which influence the institution's costs, e.g. overall policies and budgets; technological developments and changes in the behaviour of different user groups.

The institutions must be able to use the results from the cost model for strategic decision making and budgeting. It has therefore to be based on well defined accounting principles and include all costs (Full Economic Costs, FEC), including both investment costs and operating expenses, direct as well as indirect. The model must be able to measure actual baseline costs, i.e. cost based on experiences (ex-post), but the activity based approach also allows tracking cost over time and estimating future cost (ex-

[7] Charles Beagrie Limited: http://www.beagrie.com/

[8] iPRES 2009 Conference: http://www.cdlib.org/iPres/

ante), ideally within a 10 to 20 year perspective. It must therefore also be able to manage financial adjustments, such as inflation and deflation, interest and discounting.

The intention is also that the cost model should be able to be used to compare costs between the preserving institutions, and this requires that the quality of the preservation is also comparable. A basis for comparison could be obtained via certification of the preserving institutions (CRL&RLG (TRAC), 2007; DCC&DPE (DRAMBORA), 2007).

The final goal is that the user interface should be good enough that only a general knowledge of digital preservation, including information about the archive's user groups and digital collections, is needed to use the model. This must be ensured by providing the cells in the model which require input with help texts, and, where relevant, with pre-defined values. All values ought to be alterable by the user in relation to the concrete needs of the institutions.

1.3 Staffing and length of the project

The project received funding for 9 person months, divided between two staff members from DNA and one from KB. The project ran from November 2008 to December 2009.

1.4 Adjustments

In the course of the project it has been necessary to adjust the goals. While the intension was to provide a general framework for costing digital preservation, we have only been able to operationalise those parts of the model, which cover functional preservation by migration. Likewise, there has not been time for testing the model on Danish collections within ALM institutions. These adjustments are partly due to the fact that there was less help than expected to be found in the existing models. Besides, the project has been impeded due to budget stops and staff turnover.

1.5 Work methods

The project has carried out a study of the literature to uncover existing information on cost models for digital preservation. The project has furthermore done a questionnaire survey to find out which Danish experiences and studies there are in relation to cost of digital preservation. On the basis of the project's goals, the collected information and our own experience, and with point of departure in the OAIS standard, we have set up a general structure for a cost model, which we call Cost Model for Digital Preservation (CMDP). Our next step was to analyse the OAIS standard and identify those activities whose price is critical, and operationalise these by means of formulas. This work is so far only carried out for those parts of the model which relate to functional preservation through a migration strategy. Next we have tested and revised the model on the basis of the cost data collected from two migration projects (case study 1 and 2). Finally we have collected data by means of a questionnaire survey on holdings and types of materials from Danish ALM institutions to use in a comprehensive compilation of costs.

2 Existing cost models

The first task for the project was a study of the literature to find information about the general economics of preservation and, more specifically, to identify existing cost models and methods on which to build. The literature study revealed that quite a few investigations have been made of the costs of digital preservation but, as a recent report points out, these cost data are often related to specific projects, institutions or materials, and therefore difficult to transfer to other contexts (Lavoie et al., 2008, pp. 36-37).

The project identified two cost models concerned with the whole digital preservation lifecycle, namely LIFE Costing Model (McLeod, Wheatley & Ayris, 2006; Ayris et al., 2008) and Keeping Research Data Safe (KRDS1) (Beagrie, Chruszcz & Lavoie, 2008). The LIFE model was developed by the British Library and University College London, but has a generic cross-sector aim. It is inspired by a lifecycle costing methodology originally developed for paper based library collections (Stephens, 1988, 1994) and further refined for digital materials (Shenton, 2003); KRDS1 was developed by the consultancy Charles Beagrie Limited and is oriented towards the preservation of research data. The latter builds on the OAIS standard and has also been inspired by the LIFE project and by NASA Cost Estimation Tool[9]. LIFE and KRDS1 both are British and were developed with support from JISC[10], which currently also supports the further development of the models by the LIFE3 and KRDS2 projects.

These projects' resources and timescales are:

- LIFE 1, 2 and 3:

Phase	Months	External Contribution	Contributions in kind	Total
LIFE3	12	£159,969.00	£51,806.00	£211,775.00
LIFE2	18	£139,995.00	£38,551.00	£178,546.00
LIFE1	12	£102,895.00	£27,159.00	£130,054.00

- KRDS 1 and 2: 5 months £64,000 excluding VAT; and 9 months and £52,000 excluding VAT, respectively.
- NASA CET: Roughly 2FTEs per year/$250-$350 k per year, and the effort began in 2002.

The study of the literature also investigated cost models specifically directed at bit preservation and functional preservation. The Dutch National Archive has proposed a cost model, which is also expressed by means of formulas in a spreadsheet (Nationaal Archief, 2005). Furthermore, the LIFE project has developed the Generic Preservation Model (GPM) for functional preservation (Content Preservation), which also contains formulas expressed in a spreadsheet.

[9] NASA CET: http://opensource.gsfc.nasa.gov/projects/CET/CET.php

[10] Joint Information Systems Committee (JISC): http://www.jisc.ac.uk/

In the following sections the cost models mentioned above are discussed, followed by an evaluation of their usefulness in relation to this project. The models are evaluated in particular in relation to the functional model on which they are based, their use of accounting principles, their use of cost variables, as well as how the models are operationalised in spreadsheets or similar products.

2.1 LIFE Costing Model

The LIFE project has carried out a thorough literature study (Watson, 2005). The first version of the LIFE Costing Model is shown in table 1 below:

Acquisition	Ingest	Metadata	Access	Storage	Preservation
Selection	Quality assurance	Characterisation	Reference linking	Bit-stream storage	Technology watch
IPR	Deposit	Descriptive	User support		Tool cost
Licensing	Holdings update	Administrative	Access Mechanism		Metadata
Ordering & invoicing					Action
Obtaining					Quality assurance
Check-in					

Table 1 LIFE Costing Model v.1 2006.

The economic method employed in LIFE1 was evaluated in 2007 by Cloudlake Consulting Oy (Björk, 2007). The consultant remarked that the model lacked a more explicit definition of its context and that it ought to strive towards a closer adherence to the OAIS standard. The report recommended in addition that the cost used in the model should be adjusted for inflation, but that discounting should be dropped.

The LIFE model was revised in the second phase of the project, also with input from this project, and now is available in a version 2 from 2008, which is shown in table 2. The progress of the model from LIFE 1 to 2 has also been described (Wheatley, 2008).

Creation/ Purchase	Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
	Selection	Quality assurance	Repository administration	Preservation watch	Access provision
	Submission Agreement	Metadata	Storage Provision	Preservation planning	Access control
	IPR & Licensing	Deposit	Refreshment	Preservation action	User support
	Ordering and invoicing	Holdings update	Backup	Re-ingest	
	Obtaining	Reference Linking	Inspection	Disposal	
	Check-in				

Table 2 Life costing model v.2 2008.

The model calculates costs with the point of departure in the flow of the materials through the archive: Acquisition, Ingest, Bit stream Preservation, Content Preservation (functional preservation) and Access. It also has an optional stage: Creation/Purchase. The stages are further broken down and described by a series of elements and a series of optional sub-elements. The costs are calculated by adding the expenses of the individual stages together.

The LIFE Costing Model divides the costs in life cycle and non-life cycle costs, of which the latter includes management and administration, software for the preservation system, as well as inflation and discounting. It is left to the users themselves to determine whether or not non-life cycle costs should be included.

2.1.1 LIFE Generic Preservation Model (GPM)

The stage of the LIFE model, called Content Preservation, is further modeled in the Generic Preservation Model (GPM), which estimates costs for functional preservation over time. The LIFE project has produced two versions of GPM, and the formulas for the most recent version 1.3 are reproduced in table 3 below together with those parts of the LIFE Costing Model which GPM covers.

GPM is based on the premise that file formats last on an average of eight years today, but that life time is rising, so that it increases by 0,1 year per year. The expected number of preservation actions over time, is estimated from format life time. The costs related to preservation planning and quality control are moreover dependent on a file format complexity factor. File format complexity is subdivided into ten steps and ranges from 0 (low complexity) to 1 (high complexity), and is based on an estimate of the complexity of a series of formats.

LIFE2 Costing Model: Content Preservation Elements	LIFE2 Costing Model Sub-elements	Generic Preservation Model Sub-elements	GPM Formulas
Preservation Watch	1. Technology Watch 2. Monitor Institution 3. Monitor User Community 4. Monitor Producer 5. Record Planning Requirements	1. Technology Watch	Technology watch = technology watch * number of years * portion of collection
Preservation Planning	1. Preservation Planning 2. Record/Update Preservation Metadata	1. Preservation Planning	Preservation planning = preservation planning * number of years * file format complexity * portion of collection
Preservation Action	1.Integrate new preservation solution 2.Perform Preservation Action 3.QA Preservation Action 4.Record Preservation Action Metadata	1. Tool setup 2. Execute Preservation Action 3. QA	1. Tool set up costs = setup/integration of tools * expected number of preservation actions * proportion of collection 2. Execute Preservation Action = expected number of preservation actions * proportion of normalisation * number of objects * (start costs for migration /number of objects + pr. migration) 3. Quality assurance = test * number of objects * expected number of preservation actions * file format complexity
Re-ingest	1. Obtaining 2. Check-in 3. Quality Assurance 4. Metadata 5. Deposit 6. Holdings Update		
Disposal	1. Appraisal Procedure 2. Appraisal 3. Disposal Procedure 4. Disposal		

Table 3 Elements and sub-elements in the LIFE Costing Model and GPM 1.3.

After completion of the LIFE2 project GPM has been evaluated by an expert group, in which this project also participated.

2.1.2 Implementation of GPM1.3 in a spreadsheet

The LIFE project has expressed GPM 1.3 in a spreadsheet named PresModelv1-3.xls, which this project received directly from the LIFE project. The spreadsheet is very brief and simply constructed, and the user interface is incomplete. Furthermore there are many elements that the user must adjust himself, which makes it difficult for the non-specialist to use the spreadsheet.

The spreadsheet consists of four tabs named "Web Archiving Case Study", "LIFE Generic Preservation Model", "Preservation Model constants" and "Example Migration Costs". As an example of the structure of the spreadsheet, the first sheet is data from the "Web Archiving Case Study" from 2005. It consists of 45 rows and 20 columns, and provides the costs for all steps then used (Acquisition (Aq), Ingest (I), Metadata (M), Access (Ac), Storage (S) and Preservation (P), including individual sub-elements and calculated for the different types of wage levels (archivists etc.). The material is small scale (under 200 titles), and the costs for the different years are given as multiples of the average. The implementation is thus insufficiently comprehensive for use in our project.

2.1.3 Usefulness of the LIFE Costing model and GPM for the project

LIFE Costing Model is based on a well-defined function model. But it is not standardised and is generally not viewed as more detailed in its description digital preservation than the OAIS model. The widespread use of the OAIS model compared to the LIFE one also advocates for this approach. LIFE Costing Model contains the step Acquisition, which is not an immediate part of the OAIS model. Most of the steps' elements are though found in OAIS under the functional entities named Ingest or Administration. Some of the elements, e.g. Selection, are described in more detail in the LIFE Costing Model, and could be used in this project.

Both the first and second version of GPM also contain a series of formulas and variables which are interesting in relation to the operationalising functional preservation, among others, format complexity and life time.

2.2 Keeping Research Data Safe (KRDS1)

KRDS1 proposes a cost model for the preservation of research data which consists of two work tools: an activity model and a resource template, in which the costs for the individual activities are broken down. The template is based on the British accounting principle TRAC (Transparent Approach to Costing)[11], in which all costs are included (Full Economic Costs (FEC)). The report also lists cost variables distributed on general adjustments, economic adjustments and service adjustments. The activity model builds on studies of the OAIS Reference Model, LIFE Costing Model and NASA CET. It is divided into the phases 'pre-archive', 'archive' and 'support services'. Table 4 provides an overview of KRDS1.

[11] Joint Costing and Pricing Steering Group: http://www.jcpsg.ac.uk/guidance/about.htm

2.2.1 Usefulness of KRDS1 for the project

The activity model has a well-defined description of the functions. It is primarily based on OAIS, but with additions from NASA CET and LIFE, as well as from the KRDS1 project itself, e.g. the activities 'disposal' and 'first mover innovation', which it would be relevant to involve in the work of this project. A significant change in KRDS1 in relation to OAIS is that most of the functions under Administration have been moved to other functional entities. For example, the activity Preservation Action, as in the LIFE model, is added to Preservation Planning (Content Preservation). In OAIS the corresponding functions are under Administration, for example the actual preservation action is executed by Archival Information Update.

The model has an extensive and systematic description of the cost variables and dependencies which it will be useful to add to the project. And it builds on well-defined accounting principles (TRAC). KRDS1 does not include formulas for the computation of costs.

Pre-Archive Phase	**Archive Phase**								**Support Services**
Initiation	**Acquisi-tion**	**Dispo-sal**	**Ingest**	**Archive Storage**	**Preservation Planning**	**First Mover Innova-tion**	**Data Management**	**Access**	**Administration**
Creation	Selection	Trans-fer to ano-ther archive	Receive Submis-sion	Receive data from ingest	Monitor Designated Community	Develop communi-ty data standards and best practice	Admini-ster Data-base	Search and ordering	General management
	Negotiate Submis-sion Agree-ment	De-stroy	Quality Assurance	Manage Storage Hierarchy	Monitor Technology	Share develop-ment of preser-vation systems and tools	Perform Queries	Generate infor-mation package for dissemination to user	Customer accounts
	Outreach and depositor support		Generate AIP	Replace Media	Develop Preservation Standards and Strategies	Engage with vendors	Gene-rate Report	Deliver Response	Administrative support
			Generate administrative metadata	Disaster Recovery	Develop Packaging Designs and Migration Plans		Receive Data-base Updates	User support	
			Generate/update descript-tive meta-data and user docu-mentation	Error checking	Develop and monitor SLAs for outsourced preservation			New product genera-tion	
			Co-ordinate Updates	Provide copies to access	Preservation Action				
			Reference linking		Generate preservation metadata				

Table 4 Overview of the phases and activities in KRDS1 (this project's layout).

2.3 NASA Cost Estimation Tool (CET)

NASA CET is based on an assumption that it is relatively reliable to estimate future costs of data processing through an analysis of past costs. NASA CET is therefore grounded in

a series of comparative case studies (29 in all) within the framework of NASA projects, which are concerned with research observation of the earth and space. CET distinguishes between operational functions and support functions and has a 12 year time frame. CET is expressed in an Excel spreadsheet and provides a user interface through Visual Basic.

2.3.1 Usefulness of NASA CET for the project

CET does not address all the functions that a preservation institution must deal with, e.g. storage, preservation planning, software development and preservation actions. NASA CET contains a series of interesting and for this project directly useful elements: CET has a good user tool, with a built-in functionality for the computation of 'cascades', i.e. mechanisms, which make it possible to compute to what extent a change one place in the model affects other places in the model. Similarly, CET operates with an error margin of 22.9 % in its cost calculations. The method for calculating this error margin may be interesting for this project. NASA CET's empirical approach, using 29 studies to adapt the model, is an approach that CMDP would like to imitate. The possibility of this, however, relies evidently on the accessibility of such empirical data.

2.4 Digital Preservation Testbed (Nationaal Archief)

The Dutch National Archive has published a report which discusses the factors that affect costs of digital preservation. Testbed has developed a cost model based on a study of the literature, their own and others' experience, which is produced as an Excel spreadsheet, and which is used to calculate and compare costs when using various preservation strategies. The costs are divided into five indicators which are then broken down into a series of elements and sub-elements, see Table 5 below.

In the accompanying Excel spreadsheet the migration process is broken down into the following activities:

Develop preservation approach (per record type) = gather requirements + develop approach + test approach

Digital preservation activities (per batch) = Do preservation + evaluation

Cost indicators	Elementer(subelementer)
Archival system (depot/repository) Preservation systems	Physical space Hardware/software for the archival system Hardware/software for preservation systems
Personnel	Archival system Preservation systems Public services
Development/procurement of software and methods for preservation	Determine authenticity requirements Analyse authenticity requirements Design preservation approach Develop preservation approach Preservation software (parser, etc.) Viewing software Test preservation approach Document preservation approach
Performance of preservation actions (migration, migration on request, emulation)	Determine records to be preserved Construct interface with archives management system

	Incorporate systems for records management Receive records Select preservation strategy and approach Prepare records for transformation (supply metadata; repair/modify) Transformation Evaluation
Storage	Storage in archives
Other influencial factors	Public services Time between preservation actions Technology watch Supplementary storage requirements Links to management systems Volume Requirements for authenticity and reliability Preservation of the systems themselves

Table 5 Overview of cost indicators in Digital Preservation Testbed (this project's layout).

2.4.1 Usefulness of the Testbed model for the project

The breakdown of cost elements does not seem to have its point of departure in an actual functional model, but in an analysis of the Testbed system. However, the breakdown into elements and sub-elements is very detailed, and the Testbed notes a series of variables which affect the costs, and which it would be relevant to include in relation to this project. The Dutch Testbed highlights another valuable lesson that needs attention regarding the automation of a migration process: It operates with the time it takes to repair or modify records and metadata of poor quality and concludes that "This [repair] can be a slow and labour-intensive process that accounts for the majority of the costs." (Nationaal Archief, 2005, p. 11).

3 Cost Model for Digital Preservation - CMDP

The project concluded on the basis of the examination of existing cost models, that its best bet was to structure its cost model on the basis of the OAIS model. Furthermore, we concluded that it would be easier to start with OAIS from the beginning than to work in the opposite direction and try to adjust existing models such as the LIFE Costing Model to OAIS. We decided therefore that our model, which we call Cost Model for Digital Preservation (CMDP), as far as possible should be developed in accord with the OAIS structure, but benefit from the useful elements from existing models. We are aware that the OAIS model does not provide a flawless model for the development of a cost model – it covers different digital preservation areas in different levels of detail – but it has proven to be a robust basis for CMDP.

3.1 OAIS Reference Model

OAIS Reference Model is a widely used and standardised functional model, which defines concepts and processes for long term preservation systems (ISO 14721:2003). It was originally developed by The Consultative Committee for Space Data Systems, i.e. within the field of science, but has since been expanded with input from archives and libraries.

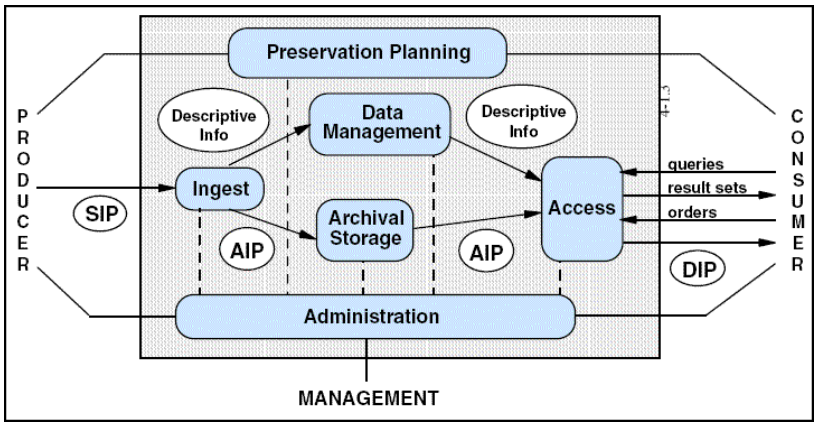

Figure 1 Shows the OAIS Model's Functional Entities, roles and data flow.

The above figure shows the OAIS model and its six central functional entities: Ingest, Archival Storage, Data Management, Administration, Preservation Planning and Access. The OAIS model consists also of the functional entity Common Services (which for clarity's sake is not included in the OAIS diagram). The diagram also shows the three roles in relation to the OAIS archive: Management, Producer and Consumer. Finally the diagram illustrates the flow of data (Information Packages, IP) through the archive in the form of Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP). Each functional entity is further broken down into a series of functions which are described in detail in the OAIS standard.

The OAIS model can be used to describe and compare preservation systems, including architectures, processes, preservation strategies and techniques, and thus in principle also the costs which are connected to the establishment and running of such systems,

and therefore it stands to reason that OAIS is an obvious foundation for a cost model. However, it is a relatively general and conceptual model. It is not aimed at conversion into concrete activities (expenses and income), which is necessary if a precise analysis of costs is to be made. Moreover, there is a difference between the degrees of detailing of the functions in the model (Eggers, 2006). Finally, the functions described are not equally costly, and here and there require more precision and thorough analysis.

3.2 Method

CMDP consists of the seven OAIS functional entities and their respective functions as well as the three roles. Thus there is a possibility that the Producer can contain the costs which correspond to the step Creation in LIFE and KRDS1.

3.2.1 Cost critical activities

To identify the cost critical activities we have analysed the function descriptions in the OAIS standards. We define cost critical activities as tasks which are estimated to take more than one person week (pw) per year to complete. Tasks which take less than one pw are added to another relevant activity, e.g. the time it takes to send a report is added to the time of the activity that it takes to produce the report. Figure 2 below shows an example of OAIS' function description of Monitor Technology, which belongs to the functional entity Preservation Planning:

The **Monitor Technology** function is responsible for tracking emerging digital technologies, information standards and computing platforms (i.e., hardware and software) to identify technologies which could cause obsolescence in the archive's computing environment and prevent access to some of the archives current holdings. This function may contain a prototyping capability for better evaluation of emerging technologies and receive *prototype requests* from Develop Preservation Strategies and Standards and from Develop Package Designs and Migration Plans. This function sends *reports*, *external data standards*, *prototype results* and *technology alerts* to Develop Preservation Strategies and Standards. It also sends *prototype results* to Develop Package Designs and Migration Plans.

Figure 2 The function description of Monitor Technology from the OAIS Preservation Planning.

Based on an analysis of the functional descriptions we have identified the flows in and out of each function and assessed which are cost critical activities. In the above case we have identified two activities as cost critical: Monitoring of Technology (emerging digital technologies, information standards and computing platforms) and Generation of reports (reports, external data standards and technology alerts). The OAIS description also says that the function "may contain a prototyping capability". We recognise that it may be necessary to make prototypes in connection with the monitoring of technology, but out of consideration for the simplicity of the model we chose to collect all the costs related to development of prototypes under the function Develop Packaging Designs and Migration Plans, which is the function that develops migration software. We have thus interpreted the individual OAIS descriptions of functions, weighed the various interests and made adjustments. Appendix 2 documents how we have done these analyses for the individual functions.

Next, we needed to breakdown the cost critical activities into measurable components in order to operationalise the formulas. The basic formula for the calculation of a cost critical activity is the actual time required to carry out the action, measured in person-weeks (pw), times the wage level (plus overhead) plus purchase (monetary value). Overhead covers indirect costs, i.e. indirect personnel, facilities, general management and administration. Wage level covers three levels: management, computer scientists and technicians. Each cost critical activity can further be regulated by different cost factors. The general structure and the breakdown methods employed in CMDP are shown in figure 1, exemplified by the functional entity Preservation Planning and its functions.

While the goal is to model the whole life cycle, the current version of the model only deals with cost of functional preservation using the digital migration strategy. For this purpose we have defined a cost critical flow between the relevant functions within the OAIS archive.

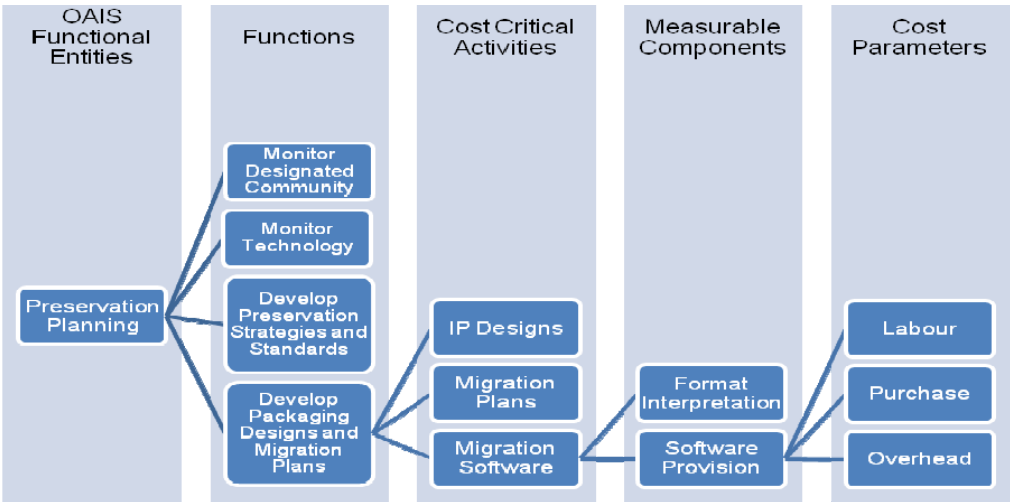

Figure 3 Overview of the structure of CMDP.

3.2.2 Cost critical activities related to functional preservation

We started with the identification of the functional entities and functions in OAIS which are activated in connection with functional preservation by migration, and defined a use case for the flow between these functions. The flow and the major cost critical activities are illustrated in Table 6.

Preservation Planning			Administration		
Monitor Designated Community/ Technology	Develop Preservation Strategies and Standards	Develop Packaging Designs and Migration Plans	Establish Standards and Policies	Manage System Configuration	Archival Information Update
Monitor					
	Develop strategies and standards				
		Develop migration package			
			Test migration package		

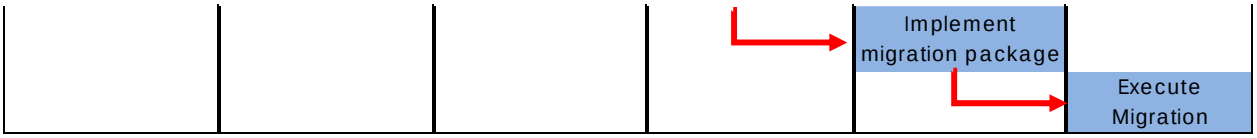

Table 6 Overview of use case for digital migration showing the cost critical activity flow through different OAIS functional entities and functions.

Note that the flow does not yet include the cost critical activities of requesting the content to be migrated from Archival Storage, nor ingesting the new Information Package (IP) version back into Archival Storage.

3.2.3 Cost factors

The complexity of digital formats and structures (objects) has a significant influence on the cost of functional preservation by migration, and it is therefore important to be able to model this factor correct. Several attempts to define this complexity have failed, and Planets' conclusion seems to be widely accepted: The notion of "digital object complexity" has been disregarded as non-objective and non-scientific (Planets, 2007). As noted the LIFE Costing Model operates with a linear scale, dividing format complexity in 10 subjectively assigned levels. We also believe that establishing differentiated complexity levels is necessary. In an attempt to tie the factor to more measurable components we have for example introduced the Format Interpretation factor, which relates the complexity to certain characteristics of the formats' documentations.

3.2.3.1 Format Interpretation

The Format Interpretation factor denotes how difficult a format is to understand. It depends on the time it takes, measured in person-weeks, to identify and read the format specifications and any other relevant documentation. Thus, it depends on the amount of documentation (number of pages); the complexity level, assigned by an estimate of whether the complexity is low, medium or high; and of the quality of the documentation, reflecting how flawed and inadequate it is (low, medium, high).

We have estimated that it takes 20 minutes on average to read and understand a page of documentation for a format with low complexity. This number is increased by 25% for a format with medium complexity, such as TIFF 6.0, and by 50% if it is has a high degree of complexity, such as GML. Regarding quality its definition is still under consideration. This parameter has not yet been implemented in the model.

Format Interpretation (pw) =
number of pages * time per page (min) * complexity (L, M, H) (* quality (L, M, H))

Table 7 shows some examples of how different formats' documentations have been evaluated in order to calculate the Format Interpretation factor. The Format Interpretation factor is used in several formulas, for example for estimating cost of Monitor Technology, Develop IP Designs, Develop Migration Plans and Migration Tools.

The Total Format Interpretation factor is the sum of an institution's format interpretation factors, each of which pertains to a specific format.

Format	Specifications and other relevant documentation	No. of pages	Complexity	Quality
TXT	ISO 10646 ISO 646	20 15	L L	H H
PDF/A 1.0	PDF/A (ISO 19005-1) PDF 1.4 (ISO 32000-1)	29 700	L H	M M
TIFF 6.0 LZW	TIFF 6.0 Baseline LZW (ISO 12639:2004)	121	M	H
GML 3.X	ISO 19136 2007 ISO 19100-serie (Open GIS) 19103 19104 19107 19108 19109 19111 19123 (understanding of xml, xml schema and Xlink assumed)	380 67 102 166 48 71 78 65	H	H

Table 7 Examples of how different formats' documentations (no. of pages, complexity and quality) have been evaluated as basis for calculating the Format Interpretation factor.

3.2.3.2 Format life expectancy and migration frequency

Formats may be migrated one at a time as they become exposed to the risk of obsolescence. However, this risk typically increases gradually. Therefore individual format migrations may be postponed in order to migrate several formats simultaneously. At the same time there are economies of scale in compiling format migration due the cost of developing IP designs, migration tools, changing work processes and system setup. Depending on the quality of the IP design, the cost of retrieving, updating and re-ingesting an information package also has important economies of scale, even though this is supposed to be fully automatic. We therefore assume that it is more likely that institutions compile format migrations to save cost.

In CMDP the frequency of migration is based on average estimated lifetime of formats, which we for simplicity have set to be 10 years. Due to variation in remaining format lifetime we estimate that migrations take place every 5 years (thereby migrating 50% of the content of the archive). In the model this factor is denoted Share of Format Replacement and thus set to 50%.

3.2.4 Application of cost factors in CMDP

The following sections describe roughly how the factors are applied in the cost critical activities under each of the affected OAIS functions.

3.2.4.1 Monitor Community and Technology

The Monitor Designated Community and Monitor Technology functions each consist of two cost critical activities, namely monitoring user community and technology, and reporting on the findings of this monitoring.

We assume that Monitor Community depends on how much influence the archive has on the production and use of formats: The more influence, the fewer costs. Monitor Technology depends on the format preserved by the archive and on those monitored.

If, for example, the archive uses preservation formats with a high degree of complexity the result is a high cost for monitoring them.

3.2.4.2 Develop Preservation Strategies and Standards

The Develop Preservation Strategies and Standards function assembles the reports received from the monitoring functions and the two cost critical activities are to develop and recommend strategies and standards (including profiles) and to provide reports. It depends on the cost of the monitoring functions.

3.2.4.3 Develop Packaging Designs and Migration Plans

The Develop Packaging Designs and Migration Plans function includes the cost critical activities of developing Information Package (IP) designs, Migration Plans and Prototypes (Software Provision). All together these 3 activities form the Migration Package:

IP Designs

IP Designs denote the structure of the container of the content in the archive. For simplicity we assume that new IP designs are required when migration is necessary. The cost of the activity is based on the Total Format Interpretation factor and the frequency of the need to create new IP designs.

Migration Plans

The activity Migration Plans includes development of general and detailed plans for migration, including test plans, community review plans and implementation plans. The cost of the activity is based on the cost of developing new IP designs and thereby, indirectly, on the cost of the Format Interpretation factor.

Prototypes

The Prototypes activity is calculated by the Software Provision factor, which comprises the cost of purchasing and/or developing prototypes and migration tools, including design, development and test. The Software Provision factor is directly dependent on the Format Interpretation factor, because the complexity of the format documentation dictates the complexity of developing migration software.

If the tools are developed the cost depends on the time (pw) it takes to develop, validate and set up the tools. To increase precision of the estimates we have divided migration tool development in three modules, namely a Reader for the source format, a Writer for the destination format and a Translator to map between the two formats. We assume that there is a base software development time of two person-days for each module and that the development time is approximately twice the Format Interpretation factor. This reflects the complexity level that the software must handle in order to read, write and interpret the formats correctly. If the source data consists of many different structures, e.g. different databases from the 1970's, the complexity is high and the tool will be correspondingly time consuming to develop. If on the other hand the tool is to be used for reading data from a text format, it will be quite simple.

If the migration tool is purchased, the required time is reduced to one third, to account for time spend on setting up and testing the purchased tool, plus the cost of the tool.

The reader and writer tool development is calculated based on the formula below, while the translator tool is calculated as the mean of the two former tools:

Development of Reader or Writer Tool (pw):
Base development + Development

Development of Translator (pw):
Base development + (Development (Reader) + Development (Writer)) / 2

Purchase of Migration tool (pw + monetary value):
Base development + Development * 33% + Purchase

The Total Software Provision factor is the sum of an institution's Software Provision factors.

Other software cost estimation tools, such as COCOMOII (Boehm, et al., 2000), use experience from similar projects and qualitative parameters or count function points for estimating the cost. This approach was however not viable for our purpose, because of lack of similar projects and uncertainty of what to develop (e.g. migration tool for an unknown destination format).

Table 8 summarises how many resources it takes to develop migration software for different likely format migrations, calculated by the formulas described above. Source formats are listed in the left column and destination formats in the first row. The numbers in italic indicate less probable migration scenarios.

Formats	TXT	TIFF	PDF/A 1.0	ISO ODT	OOXML (docx)	ISO ODS	OOXML (xlsx)	WAV	MP3	FLAC	AAC	MPEG2	MPEG 2000	SVG	JPEG 2000	TIFF
TXT		5.0	22.1	19.1	45.6	*26.3*	40.2									
TIFF	5.0		34.7	21.5	48.0	*28.7*	42.6									
PDF/A 1.0	22.1	34.7		41.1	67.6	*48.4*	62.3								28.9	
ISO ODT	19.1	21.5	41.1		64.2	Estates	58.8									
OOXML (docx)	45.6	48.0	67.6	64.2		*75.2*	89.1									
ISO ODS	*26.3*	*28.7*	*48.4*	*44.9*	*75.2*		67.0									
OOXML (xlsx)	40.2	48.0	62.3	58.8	89.1	67.0										
WAV									13.6	13.3	14.0					
MP3								13.6		13.8	14.5					
FLAC								13.3	13.8		14.2					
AAC								14.0	14.5	14.2						
MPEG2													47.7			

MPEG 2000												47.7				
SVG															30.0	25.1
JPEG 2000			28.9											30.0		12.3
TIFF														25.1	12.3	

Tabel 8 Shows how many person weeks it takes to provide software for specific migration scenarios. Numbers in italic denote less probable scenarios. Empty cells denote unlikely scenarios.

3.2.4.4 Establish Standards and Policies

The cost critical activities in Establish Standards and Policies under Administration include providing approved standards and migration goals and testing the migration package.

It depends directly on the Preservation Planning function, Develop Packaging Designs and Migration Plans, and is thus also indirectly dependent on the Format Interpretation factor.

3.2.4.5 Manage System Configuration

Manage System Configuration function includes providing reports to Develop Preservation Strategies and Standards under Preservation Planning and implementing migration packages, including tools, in the archive systems.

3.2.4.6 Archival Information Update

The Archival Information Update function consists of the cost critical activity to perform the actual migration process. In accordance with OAIS we assume that the tools and the content at this stage are flawless, which ensures an almost automatic process.

The processing time depends on the format complexity, which defines the processing speed, and the number of machines performing the migration. Based on testing we assume that the processing speed is 5 MB/s for a simple format. 25% is added if the format is of medium complexity and 50% if it has high complexity.

The model estimates the cost of man power monitoring the process to be 10% of the machine processing time. We are aware that the manual monitoring part of this process may be higher than 10 % for archive holdings that do not comply with their own preservation requirements. Work has to be done implementing a formula that calculates the cost of this eventual manual error handling of files. Alternatively a formula considering the age of yet-to-be migrated files could be implemented, so that the 'timing' (of a migration) as a cost driver mechanism would address this issue. Work also has to be done to find the correct formula concerning the processing speed of respectively small and big files.

Migration Processing (pw) = 5 MB/s * complexity (L, M, H) / number of machines * 10 %

3.2.5 Migration Cost

The Migration Cost factor is the sum of the process of interpreting a format (Format Interpretation), developing tools for a migration (Software Provision) and processing the migration (Migration Processing). The Migration Cost is calculated as shown below:

Migration Cost (pw):
Format Interpretation + Software Provision + Migration Processing

The Total Migration Cost is used to calculate the total sum of a migration action, a migration action being defined as the activities covered by the three added factors.

Table 9 lists all formulas currently applied in CMDP.

Administration (OAIS-ADM)	
4.1.1.5.B Manage System Configuration (OAIS-ADM-MSC)	
4.1.1.5.B.1	Produce MSC Reports = Base Cost 2 pw
4.1.1.5.B.2	Develop and Implement Plans for System Evolution = ROUND (Base Cost 4 pw + Migration Package*0.1;0)
4.1.1.5.C Archival Information Update (OAIS-ADM-AIU)	
4.1.1.5.C.1	Update Content = ROUND (Migration Processing;0)
4.1.1.5.E Establish Standards and Policies (OAIS-ADM-ESP)	
4.1.1.5.E.1	Test Software = ROUND (Total Software Provision*Share of Format Replacement*1.15;0)
Preservation Planning (OAIS-PP)	
4.1.1.6.A Monitor Designated Community (OAIS-PP-MDC)	
4.1.1.6.A.1	Monitor Community = ROUND (Base Cost 1 pw + Natural Logarithm (Number of Producer Groups) / Influence on Producers + Natural Logarithm (Number of Consumer Groups)/Influence on Consumers;0)
4.1.1.6.A.2	Produce MDC Reports = ROUND (Base Cost 1 pw;Monitor Community/2)
4.1.1.6.B Monitor Technology (OAIS-PP-MT)	
4.1.1.6.B.1	Monitor Technology = ROUND (Base Cost 1 pw + Total Format Interpretation;0)
4.1.1.6.B.2	Produce MT Reports = ROUND (Base Cost 1pw + 0.1*Monitor Technology)
4.1.1.6.C Develop Preservation Strategies and Standards (OAIS –PP-DPSS)	
4.1.1.6.C.1	Develop Strategies and Standards (includes Produce DPSS Reports) = ROUND (Base Cost 2 pw + 0.25*Monitor Community + 0.25*Monitor Technology;0)
4.1.1.6.C.2	Produce Recommendations on System Evolution = ROUND (Base Cost 2 pw + Monitor Technology*0.25;0)
4.1.1.6.C.3	Provide Advice on Submission Requirements = ROUND (Base Cost 2 pw + Provide MDC Reports*0.1;0)
4.1.1.6.D Develop Packaging Designs and Migration Plans (OAIS-PP-DPDMP)	
4.1.1.6.D.1	Develop IP Designs = ROUND (Base Cost 2 pw + Total Format Interpretation*Share of Format Replacement;0) Review IP Designs = ROUND (Develop IP Designs*0.15;0)
4.1.1.6.D.2	Develop Migration Plans = ROUND (Base Cost 2 pw + Natural Logarithm (Develop IP Designs);0) Review Migration Plans = ROUND (Develop Migration Plans*0.15;0)
4.1.1.6.D.3	Software Provision = ROUND (Base Cost 2 pw + Natural Logarithm (Develop IP Designs) + Total Software Provision *Share of Format Replacement*(0.25+0.125);0) Review Software = ROUND (Total Software Provision*Share of Format Replacement*1.15;0)

Tabel 9. Shows an overview of the formulas used in CMDP for digital migration.

Appendix 2 provides detailed documentation of how the OAIS functions has been broken down in cost critical activities.

4 Resume of Danish studies and experiences

In connection with the project it was investigated whether any Danish cost models and systematically documented expenses for digital preservation existed, and if so, how these could be worked into the cost model of this project. We approached the following state ALM institutions for information on Danish studies and experience: the Danish Film Institute (DFI), The National Museum (NM), The State Museum of Art (SMK), as well as SA, KB and SB. In addition contact was made with private firms, which might well have experience with cost data: Iron Mountain, Mærsk, Novo Nordisk, Scandinavian Information Audit and Recall.

Based on the answers received, the project concluded that there is no Danish cost model and only sparse systematically documented expenses for projects within digital preservation. The investigation did show that cost calculations for certain areas of digital preservation do exist, e.g. for digitalisation, bit preservation and migration.

The project considered which of these might serve as inspiration for this project's mode of **calculation** and for validation of the CMDP. So far the project has used cost data from two migration projects carried out by DNA (see the case studies below). Furthermore, the project concludes that there is relatively well-documented experience for bit preservation, and that this area does not offer any particular challenges for calculating costs. SA, SB and KB are carrying out a project on bit preservation concurrently with this one, and the work of exploiting these experiences has begun in this project.

5 Case studies

In order to test the functionality and precision of CMDP, the model was tested against cost data available from two migration projects run at DNA (case study 1 and 2).

In Denmark the DNA has the legal right to define the requirements for deliverables, with which producers must comply. These preservation requirements include specifications for data structures (IP Designs) and preservation formats. The preservation requirements are regularly revised and in principle all data in the archive updated accordingly.

5.1 Description of Case 1

The first case consists of cost data from a large migration project carried out from 2005 to 2008, where digital archives (registries and filing systems), from three different time periods, were migrated to current preservation standards:

- A-archives (1968-1998): a heterogeneous mass of data (175 MB or 3.428 files) from hierarchical databases, complying poorly with their own preservation requirements
- B-archives (1999-2000): more homogeneous data (430 MB or 9.633 files) in compliance with their own preservation requirements
- C-archives (2001-2004): data (930 MB or 12.700 files) which almost complied with the DNA's current preservation requirements

In order to make the transformation process as inexpensive, i.e. automatic, as possible, a normalised description (using XML) was made for each digital archive. Simultaneously a system was developed which could then read the normalised descriptions and transform the many variants of data structures, data types, character sets, etc., to the current preservation standard. In cases when the automatic process did not perform well, a manual process to correct errors had to be made.

A detailed registration of the incurred costs is available from the migration project distributed on work packages and tasks. On average 8 persons worked full time for three years describing the data, while 2 persons spent 2½ years developing and maintaining the system. The resources expended on the staff (subdivided in wage categories) and expenses for consultants or equipment are registered in a spreadsheet. Both the automated migration and the manual error corrections are registered, so that the numbers for both migrated data and the data types, as well as the number of files corrected manually, are available.

The automatic processing of A-archives was expensive because it frequently, roughly 80% of the time, was stopped by error messages and the need for manual correction of errors. The errors were caused by the failure of the archival materials to meet their own preservation requirements. Manual correction was necessary in 2.447 files and the cost data show that the manual corrections on the average took more or less one person-day per file, which corresponds to in all about 428 pw.

The way the migration project classified the tasks does not correspond to the way it has been done in CMDP (and therefore not in OAIS either), and thus it was necessary to map the cost data between the migration project and the CMDP. This mapping was time consuming and based on a subjective estimate. Even in so well-documented a project there is still room for doubt as to just exactly what the costs cover and how they are to be distributed in relation to CMDP. This is the case, for example, with the distribution of the aggregate project leadership expenses.

5.2 Results of Case 1

Table 10 shows the cost in person weeks (pw) of developing IP designs, Migration Plans and Prototypes (Software Provision) that is the cost of providing Migration Packages. Thus the table does not show the costs related to the monitoring functions or the processing of the migration itself. The latter is not included, because automation of the processing only reached about 20%, while the CMDP requires at least 90% automation.

The first set of columns (Case 1) gives the actual figures from case 1. The second set (CMDP) shows case 1 simulated in CMDP. The third set (CMDP-Case 1) shows the differences between case 1 and its simulation. The B&C-archives are also combined in a separate row for analytic purposes (see below). At the bottom of the table, the three activities are added up under Migration Package.

	Case 1		CMDP		CMDP - Case 1	
	pw	%	pw	%	Δ pw	%
IP Designs	**44**	**12**	**50**	**24**	**6**	**12**
A (1968-1998)	29	66	20	40	-9	-31
B (1999-2000)	15	34	16	32	1	6
C (2001-2004)	0	0	14	28	14	n.a.
B & C	15	34	30	60	15	50
Migration Plans	**150**	**42**	**39**	**19**	**-111**	**-74**
A (1968-1998)	105	70	15	38	-90	-86
B (1999-2000)	30	20	14	36	-16	-53
C (2001-2004)	15	10	10	26	-5	-33
B & C	45	30	24	62	-21	-47
Prototypes (Software Provision)	**164**	**46**	**116**	**57**	**-48**	**-29**
A (1968-1998)	101	62	48	41	-53	-52
B (1999-2000)	50	30	36	31	-14	-28
C (2001-2004)	12	7	32	28	20	62,5
B & C	62	38	68	59	6	9
Migration Package (total)	**358**	**100**	**205**	**100**	-153	**-43**
A (1968-1998)	**235**	**66**	**83**	**40**	**-152**	**-65**
B (1999-2000)	**95**	**27**	**66**	**32**	**-29**	**-31**
C (2001-2004)	**27**	**8**	**56**	**27**	**29**	**52**
B & C	122	**34**	122	**60**	**0**	**0**

Table 10 Results and comparison between case 1 and simulation of case 1 in CMDP.

Generally the comparison indicates that the CMDP underestimates the cost of the illustrated part of the Preservation Planning functions. Case 1 cost 358 pw, while the simulation outputs a cost of 205 pw – there's a deviation of 153 pw. The main reason is that the migration of A-archives was not conducted in due time, the migration should have taken place years earlier. Even though the migrated archives did not come from Producers, but from within the archive, this migration resembles a normalisation, which CMDP is not yet geared to calculate. Normalisations form part of the functional entity Ingest. The conclusion is that careful quality control of archival holdings is alpha and omega before carrying out migration, as well as that CMDP cannot be expected to estimate migration of archival holdings if the quality is too low. This latter point is also corroborated by the Dutch experience, already mentioned (2.4.1 Usefulness of the Testbed model for the project): It cost c. 333 euros for the creation of a batch of 1000 records in the pre-archive phase. In contrast once 10 years have passed and material has been transferred to an archive it may cost 10,000 euros to 'repair' a batch of 1000 records with badly created metadata (Nationaal Archief, 2005a).

If we therefore disregard the A-archives (see B&C) from the analysis and take a look at the three chunks Develop IP Designs, Migration Plans and Migration Software, we see that the CMDP is capable of estimating all of them with less deviation; albeit it pinpoints certain weaknesses pertaining to Develop IP Designs and Migration Plans (respectively 15 and 21 pw deviation), the explanation may be that the C-archives did not require IP designing, because they were almost ready for processing from the beginning. However the CMDP is designed to allocate a certain number of pw to the IP designing process. This teaches us that if data comply with the IP design at hand, the model should exclude this cost.

Regarding the Migration Plan phase, the deviation (-21 pw) is explained by the fact that the CMDP does not presently reflect the size of migration projects well enough: There is a scalability issue here, especially when the migration project uses much man-power, which requires more management. Another interesting fact is that the cost data shows that it is equally expensive to make migration plans and develop software, while the CMDP underestimates the cost of the Migration Plans step.

5.3 Description of Case 2

The second case is a current migration of 6 TB of PDF documents containing scanned property registry data to the JPEG2000 format. The PDF files are homogeneous and are each 300 MB. A detailed registration of the cost data was also available here for the test. Several of the off-the-shelf tools were evaluated, and the best purchased. It was however also necessary to develop tools in addition to purchased ones since these were inadequate to the task by themselves.

5.4 Results of Case 2

Table 3 shows the results of using the model on data from case 2 (the PDF-JPEG2000 migration). The model shows a cost of 33 pw per migration. Half of this cost in the model is due to development of migration software. In the case only 5 pw were used for the

software development. A part of the difference between the model and the case is most likely due to the model overestimating the cost of developing software migration tools; even though we have taken into account that purchasing tools only cost approximately 1/3 of in house development. Another part of the difference is most likely due to a difference in development culture between the model (based on OAIS) and the case. In the case the development was done with very little reporting and controlling. For example there were no official prototypes made for review by administration, nor any lengthy documentation.

PDF-J2K	Year	1	2	3	4	5	6	10	15	20
Monitor Designated Community		9	9	9	9	9	9	9	9	9
Monitor Technology		20	20	20	20	20	20	20	20	20
Develop Preservation Strategies and Standards		17	17	17	17	17	17	17	17	17
Develop Packaging Designs and Migration Plans		0	0	0	0	33	0	33	33	33

Table 11 Simulation of case 2 over time. Units in person-weeks.

In OAIS and therefore also in the model the function Archival Information Update performs the actual migration using the migration tools developed in Preservation Planning. The model estimates the cost of man power monitoring the process to 10% of the machine processing time. OAIS apparently assumes that once the tools have been approved by administration they are almost flawless as are the data to be migrated. In the case the migration was performed with less than 10% manpower for monitoring. One explanation for this difference is the long machine processing time in the case compared to the model. In the model we estimate a machine process speed of 2.5 MB/s for the migration of the formats on the specific hardware in use[12], but in the case the first 601 GB of data were processed at the very slow speed of 0.2 MB/s. So far 601 GB (2046 files) out of 6 TB have been migrated. One reason for the very slow process speed is probably the very large size of each file, making the process much slower than the total amount of data would indicate.

Compared to case 1 it is important to emphasize the minimal amount of manual work required to monitor the migration. The almost flawless migration process is most likely due to a high degree of compliance with the specification, i.e. very few invalid formats in the data. We estimate the compliance in the case to be above 99%. In case 1 concerning the A-archives a massive amount of manpower has been used during migration due to a very low rate of compliance (approximately 20%).

5.5 Discussion of case studies

Regarding the processing factor it is the norm to assume that the migration process is automatic. The cost of an automated process *is* quite low, but if the data to be migrated does not comply with its contemporary preservation requirements because of

[12] HW: Pentium 4 530 Prescott 3GHz, 2GB RAM, 3x7200 rpm SATA disk. Benchmarks for this machine can still be found at Tom's Hardware and other hardware sites. We have 20 machines for this type of migration, but only one was used giving these numbers.

lack of quality control (at Ingest or previous migrations), the cost of processing the data may rise exponentially due to countless hours of manual fixing. The Dutch Testbed operates with the time it takes to repair or modify records and concludes that "This [repair] can be a slow and labour-intensive process that accounts for the majority of the costs." (Nationaal Archief, 2005, p. 11). An analysis of the processing in case 1 revealed that on the average it took 1 person day to correct 1 faulty file. This example demonstrates the huge importance of compliance with preservation standards.

We have assumed that on average preservation formats will be usable for 10 years, and every 5 years a migration is performed, migrating half of the content. This is of course a very rough estimate, considering the many different types of formats and the uncertainty of technological evolution. Currently the model is not capable of varying this parameter, but we plan to enable this in future versions. For comparison The LIFE Costing Model estimates that the mean life expectancy for formats is 8 years, increasing with 0.1 year for every year that passes (McLeod et al., 2006, p. 93). Even though it is optimal not to migrate formats one by one every single year, case 1 shows that one should not wait too many years, as this becomes even more costly.

The test of the model on empirical cost data described in the case studies reveals that a very detailed and nuanced model is imperative. To exemplify this we will briefly discuss some of the most important points from running the model on test data:

A generic model should be able to handle migrations of many highly complex formats as well as a few, simple ones. It should also be able correctly to reflect the cost of projects with smaller or larger staffing. Presently, the model does not handle this well. This scalability issue does however exist on other levels too, for example concerning processing large or small files, as shown by case 2, where big PDF files process slowly, and small ones more quickly. Case 1 also demonstrates that the model cannot yet correctly calculate the cost of a migration that most of all resembles a normalisation. The model also needs more parameters to reflect that not all preconditions are fulfilled. For example in case 1 where A archives complied poorly with their own IP Design and therefore cost many pw to correct manually, while in case 2 the content complied almost fully to the IP design, and the migration was performed with minimal manual corrections. Furthermore, the model has to handle dependencies better, because no cost critical activities stand alone. Their mutual implications are difficult to account for, but highly cost sensitive. The most obvious example from case 1 was the model's difficulty of calculating the high cost of the Migration Plan phase: In our formulas this phase is dependent on the IP Design phase, but not nearly enough on the Interpretation factor (i.e. format complexity).

Regarding the degree of the precision of the CMDP we dare not yet draw any conclusions. When used for estimating future cost the precision is even more uncertain due to the challenges posed by handling the predictive element, which influences various aspects of the model. One is the life expectancy of formats, which will influence the required migration frequency. Another is estimating how much software will be

available in the future, either as open source or for purchase, and how much has to be developed. A third is estimating the complexity of future formats.

The conclusion of the above test with cost data is that CMDP manages very well, if the prerequisites for the use of the model are fulfilled. But this also means that there is a significant responsibility to inform users of the model's prerequisites.

6 Use of the model on Danish collections

The project has obtained information on a series of ALM institutions' collections, format types, systems and volume in order to use CMDP to produce a collective appraisal of the current costs of digital preservation. Due to lack of time this appraisal has not yet been carried out.

7 Conclusion

The project has carried out a study of the literature to identify cost models. Based on the literature the project identified two cost models for digital preservation (LIFE Costing Model and KRDS1) and two which target functional preservation (LIFE GPM and Testbed). The project concluded that none of the existing cost models could be directly used for the purpose of the project, but that there were many useful elements on which to build.

The project has therefore developed an independent structure for a cost model, which has been named Cost Model for Digital Preservation (CMDP). CMDP is based on the OAIS model, which provides a well-defined and standardized breakdown of the relevant activities and which encompasses all aspects of digital preservation. By including the three OAIS roles of Producer, Consumer and Management, the project has furthermore attempted to ensure that the model also takes into account the external cost factors that affect the OAIS archive, e.g. the government's policies and economic framework (Management).

The principle for the development of CMDP has been to breakdown the OAIS functions into cost critical activities and then continues to subdivide to the point of measureable components. These components are thereafter operationalised in formulas and cost factors. CMDP can at present deal with migration scenarios for a series of different preservation formats for text, pictures, sound, video, geodata and spreadsheets. To make estimating future costs more precise, it also needs to be capable of modelling migration between existing preservation formats and future (unknown) formats, as well as from unknown source formats to unknown destination formats.

The model has been tested, evaluated and adjusted iteratively on the basis of empirical cost data obtained from migration projects (case studies 1 and 2). It should be noticed that these migration projects are also a significant part of our experience, which we have used to develop the theoretical model, and CMDP has as yet not been tested on completely independent empirical data. Thus there is still another task to locate cost data for further tests.

The simulation demonstrated that the model presupposes the maintenance of a certain norm. For the one very extensive migration project (case 1) the model underestimated the large project leadership costs and the large correction costs due to a low compliance with their own preservation requirements. In the other, small scale project the model overestimated the costs of system development and the creation of new data structures.

The precision of the model is thus highly dependent on its use in a normal situation in relation to OAIS. If CMDP is to attain better precision regarding odd cases it is also necessary to make it more flexible and detailed in several areas: first and foremost for the expected life time of formats, which are currently estimated to be the same for all formats (10 years) and the migration frequency, which is set at a fixed interval (every 5th

year). There is also a need for a differentiation between new and further development of IP Designs. Nor should processing time depend solely on the complexity of formats, but also on the size of the file. Finally work must be done on to improve dimensioning the costs between large and small migration projects, as well as describing dependencies between the activities.

The Format Interpretation factor reflects the complexity of a format and is calculated by how long it takes to understand the documentation of a format. This factor appears in many of the formulas in CMDP and it is therefore particularly important to verify through test on empirical data that the way in which it is calculated is valid.

The model's user interface should not only make it possible to calculate the total costs for functional preservation, but also the costs for specific activities such as surveillance of users and technology, development of migration tools and the processing of the migration itself.

CMDP is not yet operational in relation to managing investment costs or direct and indirect running costs, which was also one of the goals of the project. Nor have mechanisms for financial adjustments yet been integrated into the model. Finally due to lack of resources it has not been possible to produce a comprehensive evaluation of the costs of digital preservation at Danish ALM institutions.

The project estimates that CMDP all in all provides a good foundation for further development and operationalising of the remaining functional entities, both with regard to assumptions, principles, methods, formulas and the user interface of the model. Furthermore it assesses that the extension of the remaining parts of the model would require fewer resources, since the principles for the method are now clarified.

A web site will as soon as possible be set up for the project from which CMDP can be downloaded and relevant documentation from the project is available: www.costmodelfordigitalpreservation.dk.

8 Future tasks

The development of a valid cost model which covers the whole of the digital preservation cycle is still the goal, and the first step, which is a natural extension of this project's results, is to extend the model to the work area covered by the OAIS functional entity Ingest, i.e. reception, handling and entering digital material into the OAIS archive.

Therefore we have proposed an extension of the project divided into three phases from 2010 -2012:

In the first phase (2010) we wish to test the theoretical model on existing data, include overhead in the model, and expand it to cover the activities of Ingest, including normalisation, which has particularly great influence on the following costs for digital preservation.

In the second phase (2011) we wish to include and test the activity Access.

In the third phase (2012) we wish to add emulation strategy under Functional preservation. The solving of this task presupposes knowledge gained in the EU project KEEP (2009-2012)[13], which aims to develop tools for emulation and analyse the prerequisites for the use of this strategy.

13 http://cordis.europa.eu/fetch?CALLER=FP7_PROJ_EN&ACTION=D&DOC=1&CAT=PROJ&QUERY=011f37a73b31:61ba:091d22f8&RCN=89496

References

Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H., & Wheatley, P. 2008. The LIFE2 Final Project Report. http://eprints.ucl.ac.uk/11758/

Beagrie, N., Chruszcz, J., & Lavoie, B. 2008. Keeping Research Data Safe. A Cost Model and Guidance for UK Universities, Copyright HEFCE 2008. http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Björk, B.C., Economic evaluation of LIFE methodology, Research report. LIFE project, 2007, p. 5-6, 18-19. http://eprints.ucl.ac.uk/7684/

Boehm, B., Abts, C., Winsor Brown, A., Chulani, S., Clark, B. K., Horowitz, E., Madachy, R., Reifer, D.J., Steece, B. 2000. Software cost estimation with COCOMO II. Englewood Cliffs, NJ:Prentice-Hall, ISBN 0-13-026692-2

Center for Research Libraries (CRL) & RLG OCLC Programs, Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, Version 1.0., 2007 http://www.dcc.ac.uk/tools/trustworthy-repositories/

Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, Blue Book, 2002, (ISO14721:2003). http://public.ccsds.org/publications/archive/650x0b1.pdf
(The OAIS standard is currently in its five years review process)

Digital Curation Centre (DCC) & DigitalPreservationEurope (DPE), Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), Version 1.0., 2007 http://www.repositoryaudit.eu/

Egger, A., Shortcomings of the Reference Model for an Open Archival Information System (OAIS). TCDL Bulletin Vol. 2(2). 2006. http://www.ieee-tcdl.org/Bulletin/v2n2/egger/egger.html

Lavoie, B. et al., 'Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation', Interrim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008. http://brtf.sdsc.edu/

McLeod, R., Wheatley, P., & Ayris, P. 2006. Lifecycle information for e-literature: full report from the LIFE project. Research report. http://eprints.ucl.ac.uk/1854/

NASA Cost Estimation Toolkit (CET). http://opensource.gsfc.nasa.gov/projects/CET/CET.php

Nationaal Archief, Digital Preservation Testbed, Cost of Digital Preservation, 2005. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf. The link to the Testbed computational model is broken (November 2009): http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185&categorie=6

Planets. 2007. Report on tool and service approach, pp. 1, 32-33. http://www.planets-project.eu/docs/reports/Planets_PA4-D1_ReportOnToolAndServiceApproach-Final_Public.pdf

Sanett, S. 2002. Toward Developing a Framework of Cost Elements for Preserving Authentic Electronic Records into Perpetuity. College & Research Libraries 63, 5, pp. 388-404.

Shenton, H., Life Cycle Collection Management, LIBER QUARTERLEY, 13, pp 294-272, ISSN 1435-5205, 2003. http://liber.library.uu.nl/publish/articles/000033/article.pdf

Stephens, A., The application of life cycle costing in libraries. British Journal of Academic Librarianship 3, pp. 82-88, 1988.

Stephens, A., The application of life cycle costing in libraries: a case study based on acquisition and retention of library materials in the British Library. IFLA journal, 1994, 20(2) pp. 130-140

Watson, J., The LIFE project research review. Mapping the landscape, riding a life cycle, 2005. http://eprints.ucl.ac.uk/1856/

Wheatley, P.: Costing the Digital Preservation Lifecycle More Effectively, iPRES2008 Conference, 2008 http://www.bl.uk/ipres2008/presentations_day1/19_Wheatley.pdf

Appendices

Appendix 1:
Kejser, U.B, Nielsen, A.B, Thirifays, A., Cost of Digital Curation: Cost of Digital Migration, Proceedings iPRES 2009 Conference.
File name: Appendix1_CMDC.doc
Power point presentation of paper at iPRES 2009 Conference.
File name: Appendix1_CMDC.ppt

Appendix 2:
Documentation of how OAIS functions have been broken down and cost critical activities identified. For the analysis we used the current public available OAIS standard from June 2009, in which changes from the 2003 edition are marked with red.
File name: Appendix2_OAISanalysis

Appendix 3:
Wordlist
File name: Appendix3_WordList.doc

Appendix 4:
Beagrie, N., External evaluation of CMDP
File name: Appendix4_DanishLife_evaluationvfinal.pdf

Appendix 5:
Questionnaire on digital material in ALM institutions
File name: Appendix5_Questionnaire for gathering information on population and anticipated growth of digital material.doc